

# Supplementary Information: Genome-wide chromatin footprinting reveals changes in replication origin architecture induced by pre-RC assembly

Jason A. Belsky<sup>1,2</sup>, Heather K. MacAlpine<sup>1</sup>, Yoav Lubelsky<sup>1†</sup>,  
Alexander J. Hartemink<sup>2,3</sup>, and David M. MacAlpine<sup>1,2\*</sup>

<sup>1</sup>Department of Pharmacology and Cancer Biology, Duke University Medical Center,  
Durham, NC 27710 USA

<sup>2</sup>Program in Computational Biology and Bioinformatics, Duke University,  
Durham, NC 27708 USA

<sup>3</sup>Department of Computer Science, Duke University, Durham, NC 27708 USA

\*To whom correspondence should be addressed: david.macalpine@duke.edu

†Present address: Department of Biological Regulation, The Weizmann Institute of  
Science, Rehovot, Israel

## Supplementary Methods

### *Construction of individual data signal tracks*

To facilitate more consistent analysis between replicates and across conditions, aligned MNase reads were subdivided into two classes: nucleosome reads (fragment width of 150–175 bp) and small fragment reads (fragment width of 25–120 bp). Within each class, the midpoint position of every sequencing read was extracted, and a 20 bp bandwidth Gaussian kernel was constructed around each position to smooth the signal tracks. Then, individual density kernels were summed together to create an aggregate signal track across each chromosome. Next, the signal tracks for each experiment were quantile-normalized to the matched small fragment or nucleosome signal from the G2 wild-type (WT) replicate 1 experiment. Finally, an overall density signal for each experimental condition was obtained by averaging the quantile-normalized signal from the two replicates. In total, eight separate signal tracks were created (three 150–175 bp nucleosome tracks and three 25–120 bp subnucleosome tracks), consisting of averaged signals across the two replicates of G1 WT, G1 *cdc6-1*, G2 WT, or G2 *orc1-161*.

A similar method was employed for the ORC and Mcm2-7 single-end ChIP-seq datasets, except that the read position was first shifted 75 bp to infer the expected midpoint location. 30 bp bandwidth Gaussian kernels were then constructed around each position and summed together to create individual data tracks, which were then similarly quantile-normalized and combined as described above.

### *Putative origin dataset*

To construct the putative origin dataset, first all 829 available ARS regions in the *S. cerevisiae* genome were obtained from OriDB (Nieduszynski *et al.* 2007). Next, the highest

scoring ACS sequence was found within each ARS region. Every position on both the Watson and Crick strands was assigned a natural log-odds ratio consisting of the probability of a sequence matching the 33 bp ACS motif (Eaton *et al.* 2010) relative to the probability the sequence originated from a 4th-order Markov model of the background sequence. Only ARS regions containing a ratio above 4 were kept for further analysis, leaving 798 ARS regions after applying this filter. Finally, these origins were centered on the highest scoring ACS position and oriented by the T-rich ACS strand.

#### *Origin efficiency determination*

An Okazaki fragment sequencing library was obtained from NCBI Accession Number SRR566701 (McGuffee *et al.* 2013) and aligned to the *S. cerevisiae* genome using the paired-end alignment strategy detailed above. An origin efficiency metric (OEM) (McGuffee *et al.* 2013) was derived around each putative ACS element by comparing the proportion of all Watson and Crick aligned Okazaki fragments ( $wl$  and  $cl$ , respectively) within a 10 kb window left of the ACS to the proportion of all Watson and Crick aligned Okazaki fragments ( $wr$  and  $cr$ , respectively) within a 10 kb window right of the ACS:  $OEM = \frac{wl}{wl+cl} - \frac{wr}{wr+cr}$ . All origins with  $OEM < 0.05$  were assigned an origin efficiency of 0.

#### *Replication timing determination*

Each putative origin was assigned either ‘early’ or ‘late’ based on evidence of replication initiation in the presence of hydroxyurea. A 5 kb window surrounding the ACS of each putative replication origin was checked for an enrichment of BrdU reads. Reads from both BrdU datasets were combined and the average number of reads expected to fall into an individual 5 kb bin was determined by assuming a random distribution of sequencing reads across the entire genome. Under these conditions, the expected

number of reads contained in any individual 5 kb bin can be modeled using a Poisson distribution. ‘Early’ putative origins were defined as any 5 kb bin with  $P < 1 \times 10^{-5}$  given this background distribution. All other origins were assigned ‘late’.

### *Individual gene locus plots*

Two-dimensional plots were constructed similar to ‘V-plots’ as described previously (Henikoff *et al.* 2011), with the x-axis representing genomic position and the y-axis corresponding to the fragment length of the sequencing read. Each sequencing read is individually plotted by first extracting the start ( $st$ ) and end ( $en$ ) coordinates along with the fragment length ( $l$ ). The sequence read is then trimmed to half its fragment length, reassigning the start and end positions to  $start = st + \frac{l}{4}$  and  $end = en - \frac{l}{4}$ , respectively. A blue rectangle is then filled with the following coordinates:  $(start, l - \frac{1}{2}), (start, l + \frac{1}{2}), (end, l - \frac{1}{2}), (end, l + \frac{1}{2})$ . Reads are individually plotted so that denser blue regions correspond to the presence of more recovered reads. Aggregate two-dimensional plots (such as those shown in Fig. ??B) are obtained by totaling the number of reads surrounding the ACS across multiple sites, with each aligned read oriented relative to the origin-specific ACS direction. Merge plots were constructed by assigning an RGB value to every square of the heatmap, where red represented the total signal from dataset one and green the signal from dataset two.

Cartoon schematics above each plot were computationally derived by finding peak positions in either the nucleosome (red) or small footprint (dark green) signal track. Nucleosome shading indicated the observed protein occupancy at a particular position, with darker red corresponding to a higher signal in the smoothed track.

### *Identifying replication origin footprints*

Putative origins were first evaluated for the presence of an ORC-dependent footprint signal in the small fragment (25–120 bp) signal track. The total signal at positions –50–150 surrounding every ACS was obtained from both the G2 WT and *orc1-161* conditions. Origins were classified as having an ORC-dependent signal if the total G2 WT signal in this region was above 100 and the fold-enrichment of G2 WT to G2 *orc1-161* was above 1.5.

For those origins without an ORC-dependent footprint, a parallel approach was conducted, except this time comparing the G1 and G2 signals. Origins were classified as having a G1-Only Footprint signal if the total G1 WT signal in this region was above 100 and the fold-enrichment of G1 WT to G2 WT was above 1.5.

### *Cell-cycle-dependent nucleosome repositioning*

G2 origin-flanking nucleosome positions were first determined for each origin in the G1 & G2 Footprint and G1-Only Footprint classes. Local maximum peaks in a 1 kb window surrounding each ACS were determined from the smoothed nucleosome signal track using the `peakDetection` function from the `nucleR` R package (Flores and Orozco 2011), requiring a minimum signal of 0.25 and ensuring that no other peak existed within 75 bp of the inferred nucleosome position. Then, each origin was assigned an upstream and/or downstream nucleosome position if the individual origin contained a nucleosome peak within 100 bp from the consensus upstream or downstream nucleosome positions (positions –90 and 148 relative to the ACS, respectively).

To find nucleosomes exhibiting cell-cycle-dependent dynamics, the G1 and G2 nucleosome signals were compared at each derived G2 nucleosome position. First, the G1 and G2 smoothed nucleosome signals were extracted  $\pm 100$  bp around the G2 nucleo-

some position. Then, each smoothed nucleosome signal was converted to a probability distribution, representing the likelihood of a nucleosome existing at a particular position. 100 random nucleosome positions were drawn from the G1 and G2 nucleosome position probability distributions, and the median G1 and G2 nucleosome positions derived from their respective samples were compared. This sampling was repeated for 500 iterations. At each origin, if at least 90% of median G1 nucleosome positions were consistently greater or less than median G2 nucleosome positions, the flanking nucleosome was considered ‘dynamic’.

Nucleosome ratio heatmaps were constructed by first extracting the G1 WT, G1 *cdc6-1*, or G2 nucleosome density  $\pm 500$  bp around each of the 398 origins in the G1 & G2 Footprint and G1-Only Footprint classes. This yielded three  $398 \times 1001$  matrices from each dataset. To normalize for MNase digestion differences, the matrices were quantile-normalized. Then, the log2 signal from G2 WT was subtracted from either the log2 G1 WT or log2 G1 *cdc6-1* samples, yielding a log2 ratio of nucleosome density at each position. Rows were ordered by the nucleosome shift classes as defined above. Genomic positions with higher ratios of G1 WT or G1 *cdc6-1* nucleosome density are colored red; likewise, higher ratios of G2 WT nucleosome density are colored green.

#### *ORC and Mcm2-7 ChIP-seq analysis*

The total ORC and Mcm2-7 ChIP-seq signal for each putative origin was determined by aggregating the smoothed signal track from positions  $-400$  to  $400$  relative to each ACS. A putative origin was assigned an ORC peak if any position contained a signal greater than 0.8; likewise, an Mcm2-7 peak was assigned if any position contained a signal greater than 0.5. The genomic coordinate with the highest signal was considered the peak position, allowing only one ORC and one Mcm2-7 peak to be assigned per origin.

389 G1 & G2 Footprint or G1-Only footprint origins (excluding 9 origins with insufficient ChIP-seq signal) were subdivided into 2 groups based on the Mcm2-7 ChIP-seq enrichment signal  $\pm 100$  bp surrounding the upstream or downstream consensus G1 nucleosome positions ( $-90$  and  $162$  relative to the ACS, respectively). To better infer the peak signal position, strand-specific ORC and Mcm2-7 ChIP-seq reads were then plotted as a merged heatmap as described above, with reads aligning to the forward strand in the red channel and reads mapping to the reverse strand in the green channel.

## References

- Eaton ML, Galani K, Kang S, Bell SP, MacAlpine DM. 2010. Conserved nucleosome positioning defines replication origins. *Genes Dev.* **24**: 748–753.
- Flores O, Orozco M. 2011. nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics* **27**: 2149–2150.
- Henikoff JG, Belsky JA, Krassovsky K, MacAlpine DM, Henikoff S. 2011. Epigenome characterization at single base-pair resolution. *Proc. Natl. Acad. Sci. U. S. A.* **108**: 18318–18323.
- MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E. 2006. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* **7**: 113.
- McGuffee SR, Smith DJ, Whitehouse I. 2013. Quantitative, genome-wide analysis of eukaryotic replication initiation and termination. *Mol. Cell* **50**: 123–135.
- Nieduszynski CA, Hiraga SI, Ak P, Benham CJ, Donaldson AD. 2007. OriDB: a DNA replication origin database. *Nucleic Acids Res.* **35**.
- Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**: 1033–1037.



## Supplementary Figure Legends

**Supplementary Figure S1.** *orc1-161* chromatin differs specifically at replication origins. (A) Nucleosome profiles surrounding yeast transcription start sites (TSS) (Xu *et al.* 2009) for WT and *orc1-161* digested chromatin. The small fragment (25–120 bp) footprint occupancy for WT and *orc1-161* at (B) Abf1 binding sites (MacIsaac *et al.* 2006) and (C) replication origins.

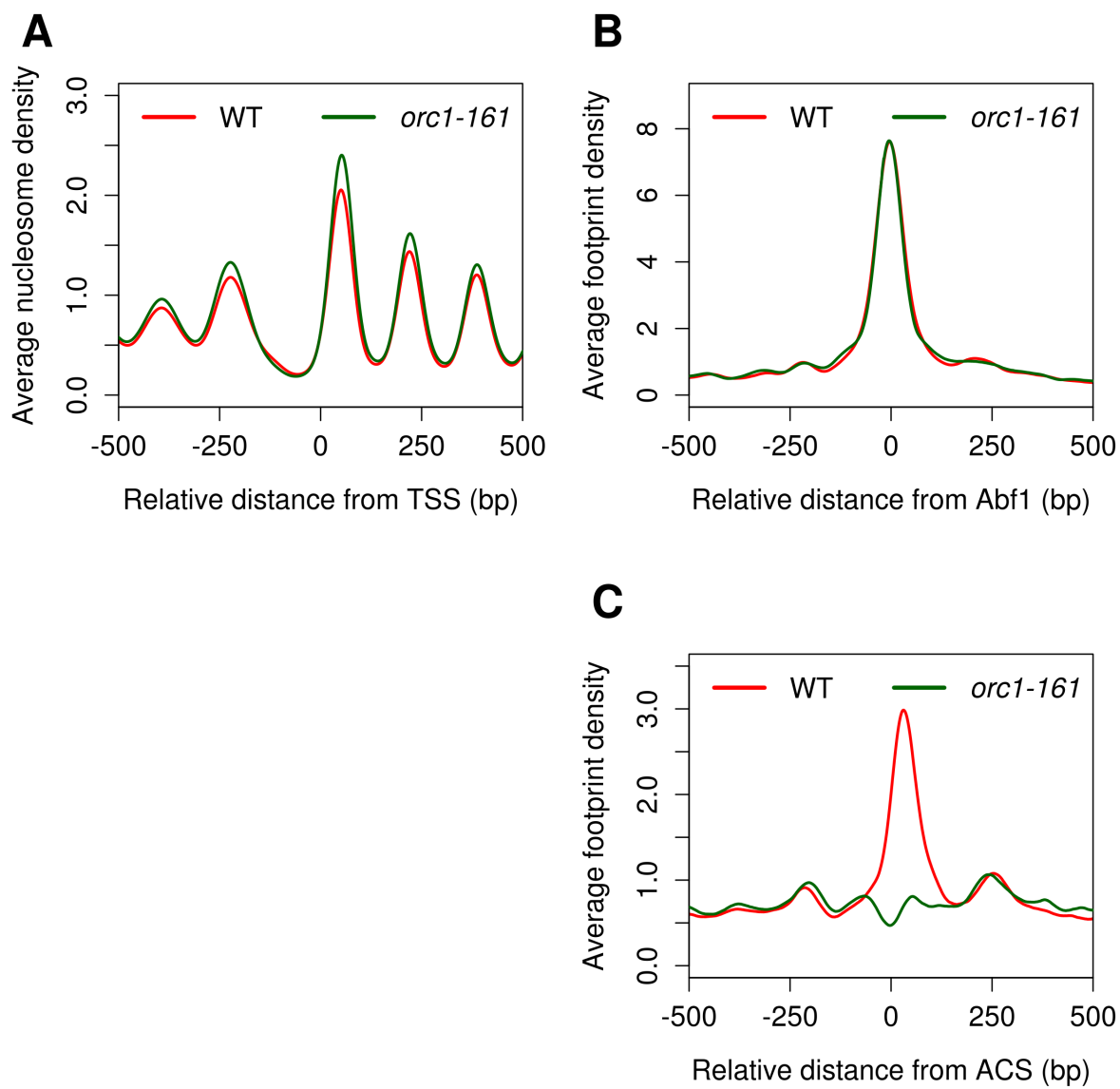
**Supplementary Figure S2.** Replication origins increase protein occupancy in G1. The G1 small fragment (25–120 bp) occupancy for G1 WT (red), G1 *cdc6-1* (purple), and G2 WT (green) for (A) replication origins and (B) Abf1 binding sites (MacIsaac *et al.* 2006). (C) ORC (top) and Mcm2-7 (bottom) occupancy from WT and *cdc6-1* strains were determined by ChIP in G1 and evaluated by quantitative-PCR. Fold ChIP enrichment was normalized to *ARS1* WT.

**Supplementary Figure S3.** ACS sequence is not predictive of occupancy differences between the G1 & G2 Footprint and G1-Only Footprint classes. Position-weight matrices were derived from the ACS sequences at (A) G1 & G2 Footprint, (B) G1-Only Footprint, or (C) No Footprint replication origins.

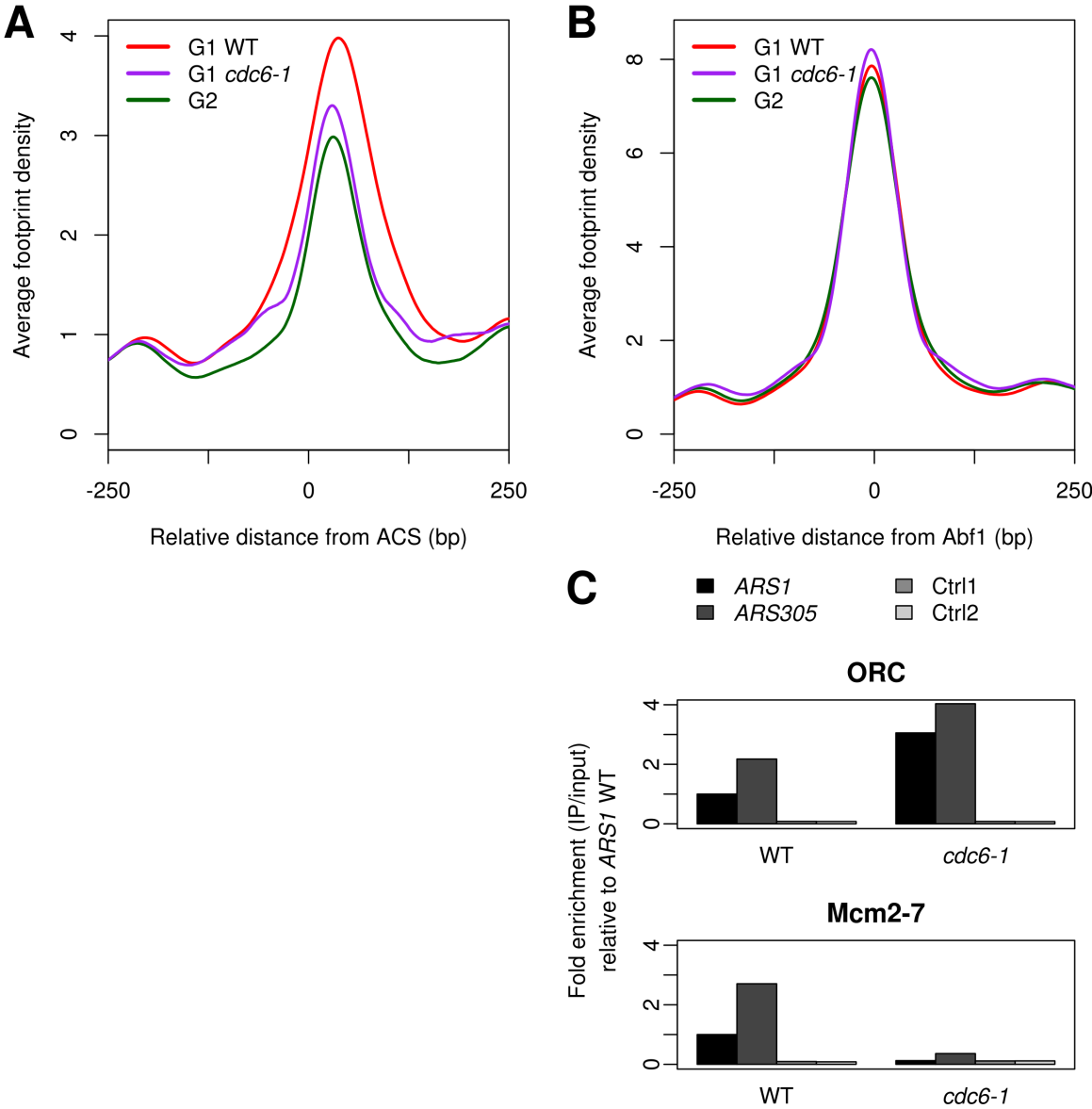
**Supplementary Figure S4.** Aggregate cell cycle nucleosome densities for WT and *cdc6-1*. (A) Average G1 WT and G2 WT nucleosome densities were generated for each nucleosome class identified in Fig. 4. (B) Average G1 *cdc6-1* and G2 WT nucleosome densities for the same origin classes as in (A). (C) Distribution of G1 Mcm2-7 ChIP-seq enrichment across each nucleosome class. Mcm2-7 ChIP-seq signal was significantly less in the static class relative to the dynamic nucleosome classes (Wilcoxon test, upstream vs. static,  $P < 5.99 \times 10^{-4}$ ; downstream vs. static,  $P < 7.95 \times 10^{-7}$ ). The upstream and downstream nucleosome classes did not significantly differ in Mcm2-7 occupancy ( $P < 0.136$ ).

**Supplementary Figure S5.** Apparent simultaneous DNA occupancy by Mcm2-7 and nucleosome cannot be explained by stochastic competition or Mcm2-7 protection of a ~150 bp nucleosome-sized fragment. Average nucleosome densities for the (A) upstream and (B) downstream Mcm2-7 loading classes from Fig. 5C show no substantial cell-cycle-dependent nucleosome occupancy differences. Nucleosome density is not greatly lost between G1 and G2 at the Mcm2-7 loading position, nor is nucleosome occupancy significantly different between either flanking nucleosome. (C) Fragment length distribution of origin-flanking regions for upstream (left) and downstream (right) Mcm2-7 occupancy classes. (D) Origin activation and efficiency for each Mcm2-7 occupancy class. No origin efficiency differences (Wilcoxon test,  $P < 0.4664$ ) or activation time differences (Chi-square test,  $P < 0.055$ ) were detected.

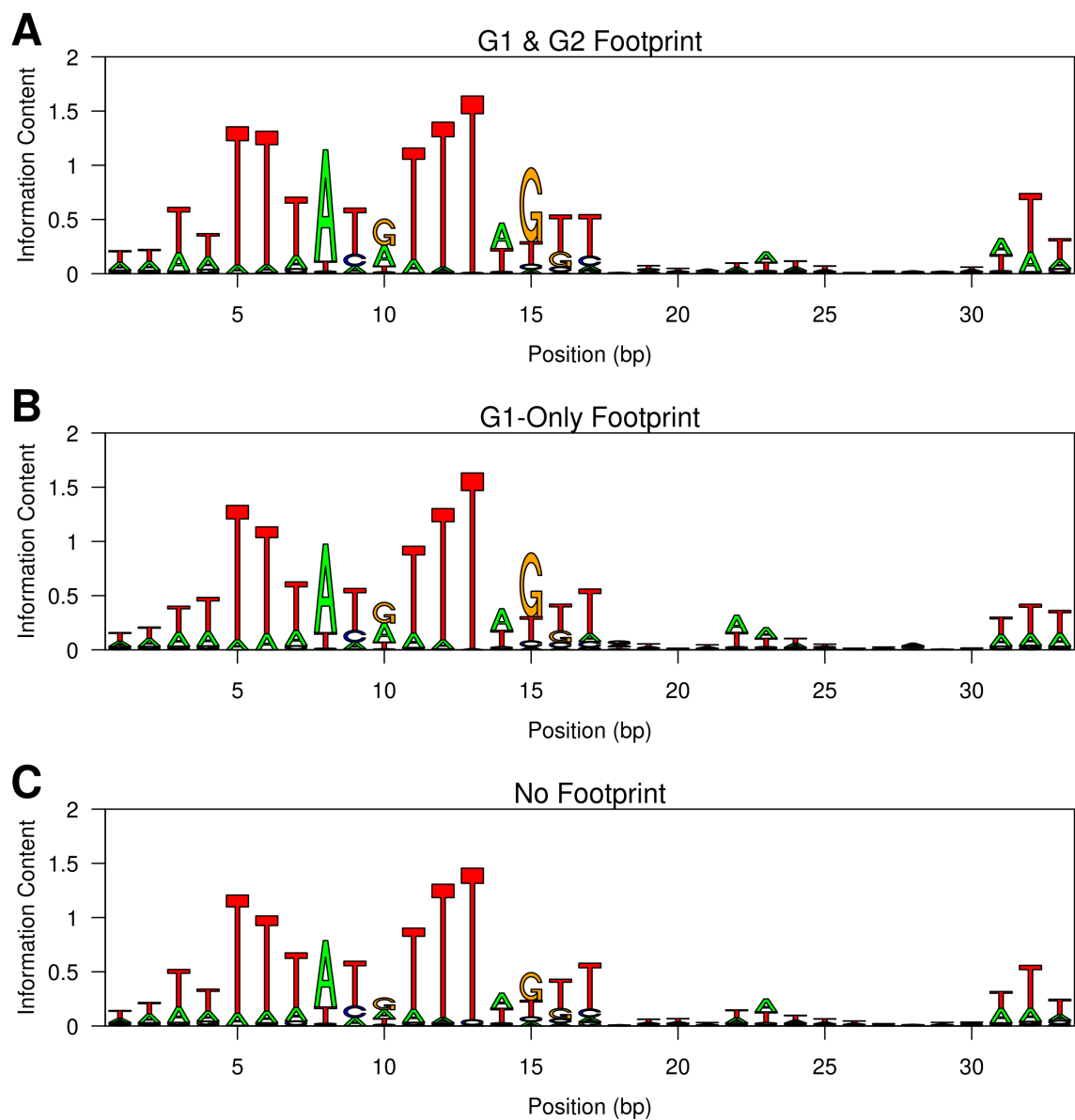
**Supplementary Figure S6.** Abf1p ChIP-seq shows protein occupancy in the middle of the NFR. Analysis of chromatin fragmentation patterns resulting from Abf1p ChIP-seq. A heatmap representation of forward (red) and reverse (green) read distributions at each Abf1 binding site (MacIsaac *et al.* 2006). Top panel represents the average forward and reverse ChIP-seq densities, and most probable fragmentation boundaries are designated by red and green dotted lines, respectively. The average nucleosome density (blue) for each class is superimposed, and inferred nucleosome positions are depicted as red shaded ovals.



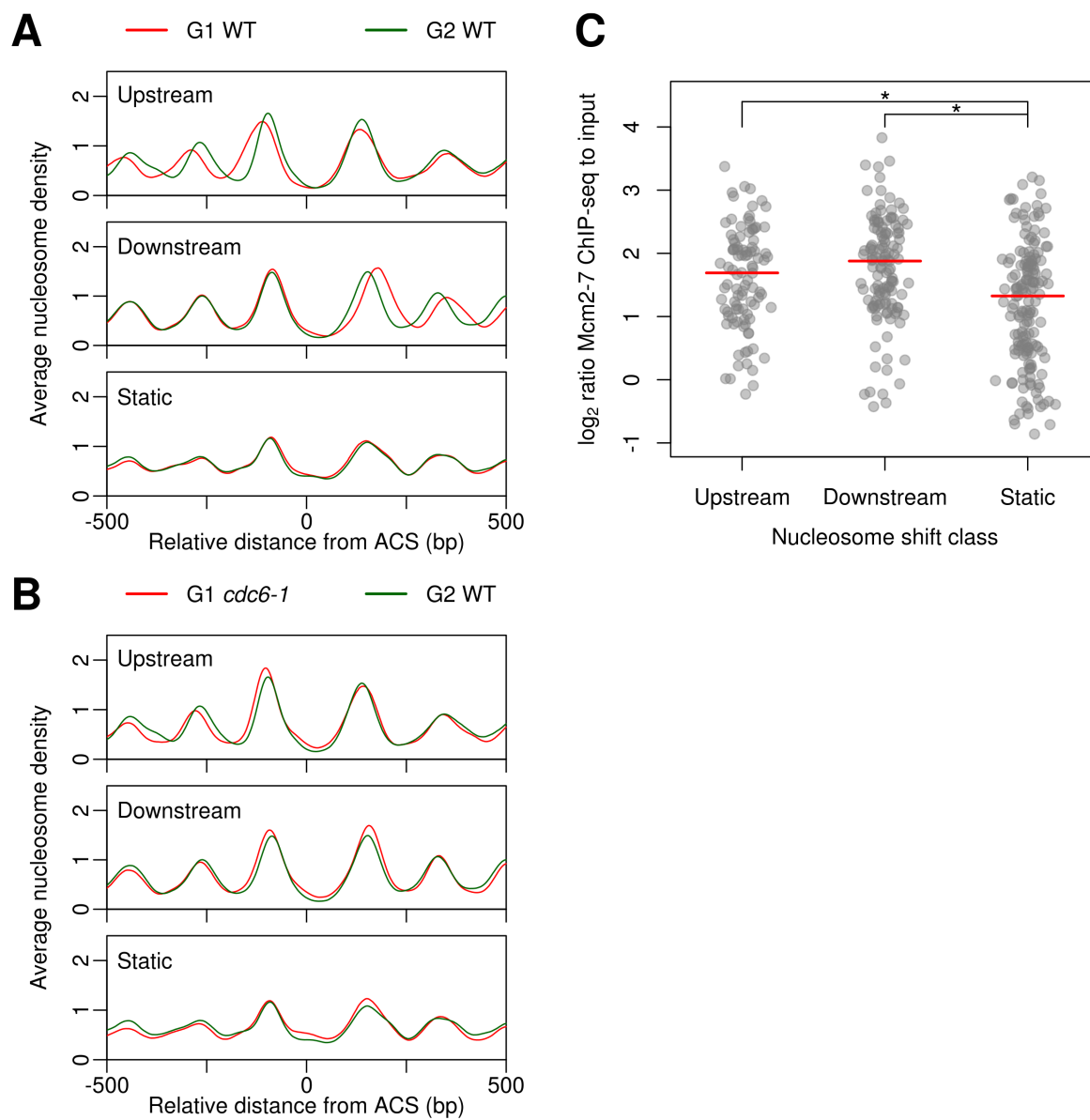
**Supplementary Figure S1.** *orc1-161* chromatin differs specifically at replication origins.



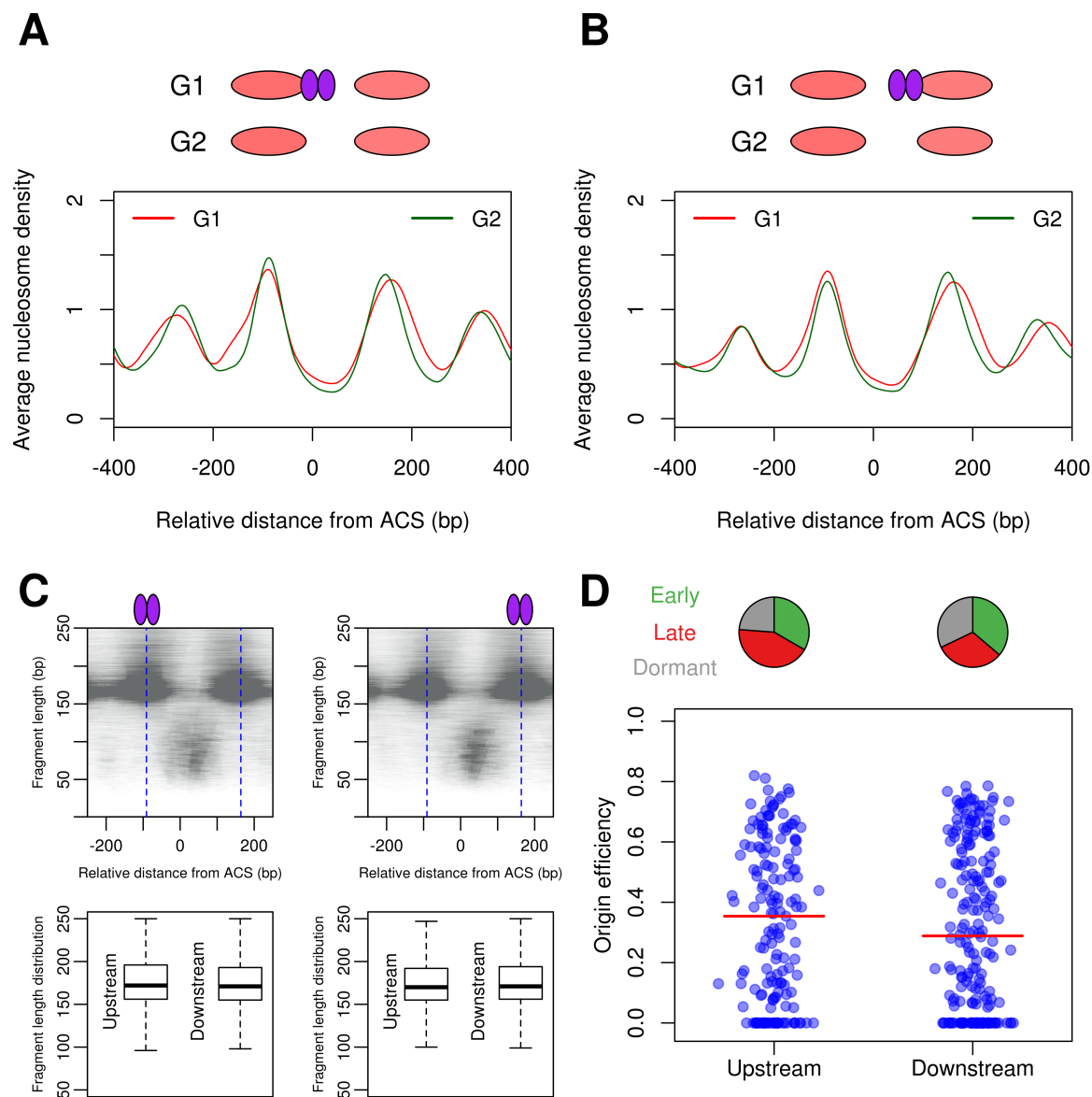
**Supplementary Figure S2.** Replication origins increase protein occupancy in G1.



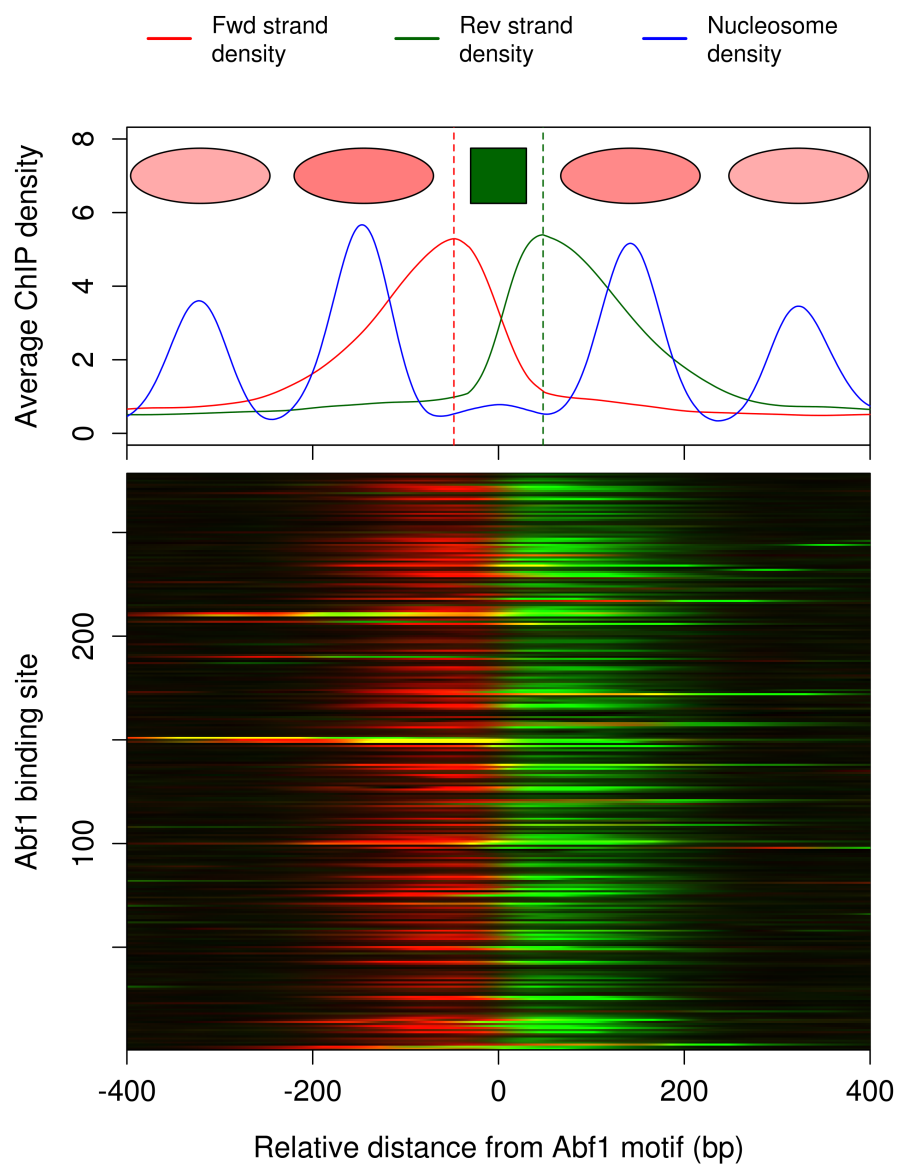
**Supplementary Figure S3.** ACS sequence is not predictive of occupancy differences between the G1 & G2 Footprint and G1-Only Footprint classes.



**Supplementary Figure S4.** Aggregate cell cycle nucleosome densities for WT and *cdc6-1*.



**Supplementary Figure S5.** Apparent simultaneous DNA occupancy by Mcm2-7 and nucleosome cannot be explained by stochastic competition or Mcm2-7 protection of a ~150 bp nucleosome-sized fragment.



**Supplementary Figure S6.** Abf1p ChIP-seq shows protein occupancy in the middle of the NFR.